DOCUMENT RESUME

ED 459 830 IR 058 366

AUTHOR Cooper, James W.; Viswanathan, Mahesh; Byron, Donna; Chan, Margaret

TITLE Building Searchable Collections of Enterprise Speech Data.

PUB DATE 2001-06-00

NOTE 11p.; In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (1st, Roanoke, Virginia, June 24-28, 2001). For entire proceedings, see IR 058 348. Figures may not reproduce well.

AVAILABLE FROM Association for Computing Machinery, 1515 Broadway, New York NY 10036. Tel: 800-342-6626 (Toll Free); Tel: 212-626-0500; e-mail: acmhelp@acm.org. For full text: http://www1.acm.org/pubs/contents/proceedings/dl/379437/.

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Databases; Information Processing; Information Retrieval; Information Seeking; Internet; *Online Searching; World Wide Web

IDENTIFIERS Data Management; *Data Mining; *Speech Recognition

ABSTRACT

The study has applied speech recognition and text-mining technologies to a set of recorded outbound marketing calls and analyzed the results. Since speaker-independent speech recognition technology results in a significantly lower recognition rate than that found when the recognizer is trained for a particular speaker, a number of post-processing algorithms was applied to the output of the recognizer to render it suitable for the Textract text mining system. The call transcripts were indexed using a search engine and Textract and associated Java technologies were used to place the relevant terms for each document in a relational database. Following a search query, a thumbnail display of the results of each call was generated with the salient terms highlighted. These results are illustrated and their utility is discussed. Results of these experiments were taken and this analysis was continued on a set of talks and presentations. A distinct document genre is described, based on the note-taking concept of document content, and a significant new method is proposed for measuring speech recognition accuracy. This procedure is generally relevant to the problem of capturing meetings and talks and providing a searchable index of these presentations on the Web. (Contains 19 references.) (Author/AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 459 830

Building Searchable Collections of Enterprise Speech Data

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D.Cotton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

By: James W. Cooper, Mahesh Viswanathan,
Donna Byron & Margaret Chan

IR058366

BEST COPY AVAILABLE

Building Searchable Collections of Enterprise Speech Data

James W. Cooper, Mahesh Viswanathan
IBM T J Watson Research Center
PO Box 704, 208
Yorktown Heights, NY 10598
001 914-784-7285, 001 914-945-1754
jwcnmr, mashehv@watson.ibm.com

Donna Byron
University of Rochester
Rochester, NY
dbyron@cs.rochester.edu

Margaret Chan
Columbia University
New York, NY
Maggie@scientist.com

ABSTRACT

We have applied speech recognition and text-mining technologies to a set of recorded outbound marketing calls and analyzed the results. Since speaker-independent speech recognition technology results in a significantly lower recognition rate than that found when the recognizer is trained for a particular speaker, we applied a number of post-processing algorithms to the output of the recognizer to render it suitable for the Textract text mining system.

We indexed the call transcripts using a search engine and used Textract and associated Java technologies to place the relevant terms for each document in a relational database. Following a search query, we generated a thumbnail display of the results of each call with the salient terms highlighted. We illustrate these results and discuss their utility. We took the results of these experiments and continued this analysis on a set of talks and presentations.

We describe a distinct document genre based on the note-taking concept of document content, and propose a significant new method for measuring speech recognition accuracy. This procedure is generally relevant to the problem of capturing meetings and talks and providing a searchable index of these presentations on the web.

Categories and Subject Descriptors

Sound information, Information repositories, Speech information, Evaluation methods, Human-computer interaction, Interface design, Visualization, Concept representation, Document genres, Markup schemes, Metadata, Information retrieval, Multimedia retrieval, Text retrieval.

General Terms

Algorithms, Management, Measurement, Design, Experimentation, Human Factors.

Keywords

Speech analysis, Speech retrieval, Text mining, Search, Document display.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. INTRODUCTION

The problem of finding important and relevant documents in an online document collection becomes increasingly difficult as documents proliferate. Our group has previously described the technique of Prompted Query Refinement [7, 8] to assist users in focusing or directing their queries more effectively. However, even after a query has been refined, the problem of having to read too many documents still remains.

We have also previously reported the details of the "Avocado" summarization system we developed for producing rapid displays of the most salient sentences in a document. [13]

Users would prefer to read or browse through only those documents returned by a search engine that are important to the area they are investigating. We have previously described document retrieval systems that can utilize a set of relatively easily derivable numerical parameters to predict which documents will be of most interest to the user. [5]

We now report applying these techniques to data from a speech recognition engine. Since this technology assumes that input is well-edited text, such as articles or news stories, performing this mining on the output of the speech engine represents a new and somewhat complex challenge.

We first describe how we obtained the initial dataset we studied, and then describe the processing necessary to obtain transcripts of these data. Then, we describe post-processing of the speech transcripts and how text mining was carried out on the transcripts. We describe two client server systems and user interfaces we developed for representing these data, and discuss the advantages and limitations of the speech mining techniques. We propose a simple technique for evaluating the accuracy of speech transcripts of these informal conversations.

Finally, we describe two experiments in indexing and summarizing consultant reports and technical talks, and compare our results with those found by human listeners. We describe some of the most promising applications of this system.

2. BACKGROUND

Finding documents in a collection is a well-known problem and has been addressed by any number of commercial search engine products, including Verity, IBM Intelligent Miner for Text and Alta-Vista.

There have been a number of approaches to solving document retrieval problems in recent years. For example, Fowler [10] has described a multi-window document interface where you can drag terms into search windows and see relationships between terms in

a graphical environment. Relevance feedback was utilized by Buckley [3] and Xu and Croft, [19] who also utilized local context analysis using the most frequent 50 terms and 10 two-word phrases from the top ranked documents to perform query expansion. Schatz *et al* [16] describe a multi-window interface that offers users access to a variety of published thesauri and term co-occurrence data.

2.1 The Talent Toolkit

Our group at IBM has developed a number of technologies to address these problems. In thisproject, we utilized the suite of text analysis tools collectively known as Talent (Text Analysis and Language Engineering Tools) for analyzing all the documents in the collection. The portions of TALENT relevant to these experiments are described in the following sections.

2.2 Textract

The primary tool for analyzing this collection is **Textract**, itself a chain of tools for recognizing multi-word terms and proper names. Textract reduces related forms of a term to a single *canonical form* that it can then use in computing term occurrence statistics more accurately. In addition, it recognizes abbreviations and finds the canonical forms of the words they stand for and aggregates these terms into a vocabulary for the entire collection, and for each document, keeping both document and collection-level statistics on these terms.

Each term is given a collection-level importance ranking called the IQ or Information Quotient [5, 7]. IQ is effectively a measure of the document selectivity of a particular term: a term that appears in "clumps" in only a few documents is highly selective and has a high IQ. On the other hand, a term that is evenly distributed through many documents is far less selective and has a low IQ. IQ is measured on a scale of 0 to 100, where a value of X means that X% of the vocabulary items in the collection have a lower IQ. Two of the major outputs of Textract are the IQ and collection statistics for each of these canonical terms, and tables of the terms found in each document.

2.3 Context Thesaurus

We have previously described the context thesaurus [7, 8] It is computed from a concordance of all the sentences and occurrences of major terms in those sentences. It is an information retrieval (IR) index of the full text of the sentences surrounding these terms and thus provides a convenient way for a free text query to return terms that commonly co-occur with the query phrase. It also provides an entry point into the collection of terms that actually have been found in the collection, rather than terms that might be predicted *a priori* or using standard dictionaries and thesauri. It is similar to and was inspired by the Phrase-finder [19].

2.4 Named and Unnamed Relations

The Textract system also produces tables of discovered named and unnamed relations. Unnamed relations are strong bi-directional relations between terms which not only co-occur but occur together frequently in the collection. These terms are recognized from the document and term statistics gathered by Textract and by the relative locations of the terms in the document. Named relations [2] are derived by a shallow parsing of the sentences in each document, recognizing over 20 common English patterns which show a named relation between two terms.

3. THE TECHNICAL CHALLENGE

In phase one of this project we worked with a customer in a financial services organization to investigate the effectiveness of analyzing outbound marketing telephone calls with several objectives in mind. It was suggested in preliminary discussions that it might be possible to mine additional information about customers that would be useful in identifying additional financial products that might be of value to them. For example, one could imagine key concepts like "college-age" or "retirement" being used to help assess a customer's financial strategy. Additionally, initial discussion suggested that we might be able to profile the performance of a successful sales call or of a successful salesman by comparing these transcripts with sales success data.

While we began this particular study because of connections to customers, we feel that the overall approach is broadly applicable to any number of libraries of talks and other speech data, such as phone messages, conference calls, and meetings.

Over the course of the project we modified these objectives to ones that while more modest, provided significant benefits to the customer in analyzing the calls.

Data were captured by analog recording of outbound marketing calls for 16 salesmen over a 4-week period. The data were then manually transcribed using a transcription service and used as a model to evaluate the accuracy of speech recognition when applied to these recordings.

The speech recognition system we used was the Large Vocabulary Telephone Transcription system (LVTT), developed from IBM's large vocabulary continuous speech recognition system (LVCSR) [4]. Use of this system in indexing broadcast news has been discussed by Dharanipragada and Roukos [9] and by Viswanathan *et al* [17].

The LVTT system is a large vocabulary speaker-independent system for recognizing and transcribing speech from telephone calls. For this project, the vocabulary was enhanced using terms found on the financial vendor's web site.

Initial results using analog recordings were not promising, and the recordings were repeated using a digital recording system. This system recorded 16 marketing personnel during a single month. To avoid excessive disruption to the customer's business, eight salesmen were recorded for two weeks and the other eight during the remaining two weeks.

This resulted in about 11,000 logical phone call units, including random misfires of the recording system, "he's not here" calls and some interspersed business and personal calls. All of these were processed using the LVTT system and provided as text files for further analysis.
Of these 11,000 calls, we excluded all calls with less than 3K of text as not containing any useful information. This reduced the actual number of calls we analyzed to 523.

4. ANALYSIS OF SPEECH RECOGNIZED DATA

Speech recognized data from telephone calls presents some unique challenges compared to, for example, the high quality PC-based dictation systems (such as IBM's ViaVoice) now available. Not only must the speech recognition be speaker-independent, but it must also deal with a wide variety of accents for both the marketing people and the customers, and significantly reduced audio quality. In this case, the callers and most of the customers

called had a wide variety of difficult regional accents, in addition to any number of foreign accents.

Finally, and most significant, telephone conversation is informal speech, consisting of phrases, fragments, interruptions and slang expressions not normally found in formal writing. Thus, the predictive model that speech recognition engines use to recognize which words are likely to come next is much more likely to fail.

Speech recognition systems are built on two models: a language model and an acoustic model. The acoustic model for telephone transcription can help mitigate the reduced frequency spread in the resulting recording. The language model is built on some millions of words found in general writing. It can be enhanced by including domain terms for the area being discussed, in this case, for the financial industry.

However, even with this additional enhancement, the quality of speech recognition is at best 50% of the words in the telephone conversations, and in problematic cases significantly worse. It is these data that we then sought to process and mine into information useful to our research and eventually to our commercial partner.

4.1 Analysis of Raw Recognition Results

As we indicated earlier, the word error rate in these samples of speech recognized data was no more than 50%. Further, the transcripts were not divided either by speaker or even by sentence. A typically problematic transcript fragment is shown below:

that has sweat what you have a minus for the one year before that you you look have all along are right you feel that has performed for you right-now one term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man yeah time you want my name and then I'm-sorry a blue-chip fund number's one because the middle bond fund over ten year period has returned an IRA over a five year period of five point eight are

While there is clearly information in this fragment, we can see that it is going to be extremely difficult for text mining technologies to pull out useful concepts from such data.

Accordingly we developed a number of algorithms to process these data further before submitting them to the Textract text mining and search engine indexing processes. Text mining assumes well-edited text, such as news articles or technical reports, rather than informal conversation, inaccurately recorded.

Much of the post processing analysis we performed on these call transcripts was driven by Textract's requirements of well-edited text in sentences and paragraphs. Textract uses these boundaries to decide whether it can form a multiword term between adjacent word tokens and how high the level of mutual information should be in determining co-occurring terms.

4.2 Timing Information

We first used the timing information in the raw speech data to insert periods and paragraph breaks in the text stream. While the speech recognition engine provided estimates of these points, we were able to fine-tune this process by applying our own empirically derived parameters. In this suite of calls, we replaced pauses of between 0.80 seconds and 1.19 seconds with a sentence break. Specifically, we added a period, two blanks and capitalized the following word.

We replaced pauses of 1.2 seconds or more with a new paragraph, by adding a period, two blank lines and a capital letter to the next word. Paragraph boundaries were important in this analysis because speaker separation information was not available in this research version of the voice recognition engine, and in mining text for related terms, paragraph boundaries provide a break between sentences that reduces the strength of the computed relationships between terms.

The speech engine provided silence information as a series of "silence tokens," where each one was assigned a duration. Frequently, there would be several sequential silence tokens, presumably separated by non-speech sounds. When this occurred, we summed the silence tokens to a single token that we used to determine whether to insert punctuation.

4.3 Word Certainty

The LVTT speech engine provided us with estimates of the certainty it had recognized a word correctly. It also provided a series of alternate choices that it had considered less likely.

Our initial theory was that if we examined the words and alternates as a continuous matrix, we might be able to recognize multi-word terms among the words and alternate choices to improve the quality of recognition. This is similar to the procedure suggested in TREC-8 by Johnson [11]. However, our analysis showed that considering these alternate choices and looking for phrases among them did not result in any improvement at all. In fact, we did not discover any cases where choosing word alternates would improve recognition. This is probably not surprising, because the predictive speech models used in recognition already take these facts into account.

We also investigated document expansion using a parallel corpus in a manner similar to that suggested by Woodland [18]. We indexed a number of well-formed business documents describing the customer's business and products. Then we queried the index using each of the transcribed calls and augmented the transcribed calls with the few documents returned from this query. The purpose of this exercise was to overcome word recognition errors to try to improve recall. We found a very small improvement and decided that it was not sufficient to warrant the large amount of extra computation to achieve it.

We did find the certainty figures useful in the call analysis in another significant way. If the speech engine indicated that a word was recognized with low certainty we considered whether eliminating the word would provide a more useful transcript. This became important because the speech engine tended to insert proper nouns for words it recognized with low certainty and these nouns were frequently incorrect. When the Textract text-mining system is run on such text, these proper nouns are recognized as salient when they in fact should not have been found at all.

Our initial impulse was to remove these low confidence terms from the transcripts entirely, but this would lead to the text-mining system forming multi-word terms across these boundaries when it should not have been able to do so. So instead, we replaced each occurrence of a low confidence terms with the letter "z." These non-word tokens prevented the formation of spurious multiwords without significantly reducing the clarity of the transcript.

For example, in one call we found the sentence fragment:

Over a five year period has returned an IRA...

where the final word "IRA" was of low certainty. Our algorithm converted that to

Over a five year period has returned an z...

which removes the incorrectly recognized token "IRA" that was not in fact part of the originally spoken sentence, but would lead to a wildly misleading summary of the document. (In fact the correct phrase was "returned a 6 and one-half percent...")

The speech engine also produced tokens for non-word utterances such as "uh," "um" and <smack> which we removed entirely. Removing these was actually quite necessary, since they often interfered with the formation of multi-word terms, so that *bond <uh> funds* was reduced to *bond funds*.

5. USING LINGUISTIC CUES

In addition to our reanalysis of the data provided by the speech engine, we also realized that there are some English language cues we can use to improve our recognition of sentence boundaries. There are a number of common English words and phrases which are used exclusively or primarily to start sentences, such as *Yes, OK, Well, Incidentally, Finally* and so forth. In consultation with other linguists in our group, we tabulated a list of these words and phrases and then used them to further process our transcripts. Whenever we found such words or phrases, we inserted a period, two spaces and capitalized the token we found. In these two-way conversations, we found that in most cases, we could insert a paragraph boundary as well. Applying all of these techniques to the initial text we showed above, gives the somewhat more coherent transcript below.

Yeah, that has sweat what.
You have a minus for the one year before that you look have all along are.
Right, you feel that has performed for you right now one Term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man.
Yeah, time you want my name and then I'm sorry a blue chip fund number's one because the middle bond fund over ten year period has returned an z. Over a five year period of five point eight.

6. TRANSCRIPT ANALYSIS

Once we performed all these analyses and post-processing techniques, we still had very poor and confusing transcripts to deal with. Considering the technological barriers we had to overcome, this is not entirely a surprise. However, the question then arose as to whether there was any value at all in such transcripts. It turns out that there definitely is a great deal of information in these documents once we overcome the idea that we are producing transcripts of conversations.

If instead we consider the idea that we are actually producing notes on a meeting to help in summarizing what transpired, we find that there is real value to be extracted even from these noisy data. For example, even from the intentionally vague selection quoted above, we find that we can extract the concepts

- Minus for the one year
- Average top ten
- Blue chip fund
- Middle bond fund
- Ten year period
- Five year period

From those concepts we can begin to outline the nature of the conversation, even without accurate speech recognition.

6.1 Indexing the Transcript Files

Once the transcripts were processed as we described above, we analyzed them using Textract and indexed them using a standard search engine indexing system.

We found that in these unstructured conversations, Textract was not able to form any significant named relations and only a very few unnamed relations, so we did not use this information in constructing our retrieval system.

As we have described previously [13], we used a set of Java programs to analyze the Textract output and load it into a database. From this database, we can easily ask for the most salient terms in any document, and can even restrict the query to the most salient multiword terms in the document.

For example, for the complete conversation we excerpted above, the most salient terms are shown in Table 1. Textract categorizes terms into eight types. In this table and in similar queries in this research, we excluded terms assigned to the Unknown Name (Uname) and Unknown Term (Uterm) categories, which produced a large quantity of fairly uninformative terms like "room."

Table 1- Terms discovered in a single conversation using Textract

Term	IQ
Middle bond fund	85
Chemical bond fund	85
Negative point	85
Strategic income	50
Ginny Mae	47
Percent return	40
Year period	24
Core bond	16
Tax exempt	6
Capital gain	3

7. A CLIENT-SERVER CALL QUERY SYSTEM

After constructing these indexes, we were able to construct a Java-based client-server search system.

The server consists of a search engine index, and a document and terms database. Here the search engine was initially IBM's TSE search engine, later replaced with IBM's GTR search engine, and the database was DB2. The system is driven using a Java RMI server and communicates with a Java RMI client running in a web browser, using the Java plug-in. This system is illustrated in Figure 1.

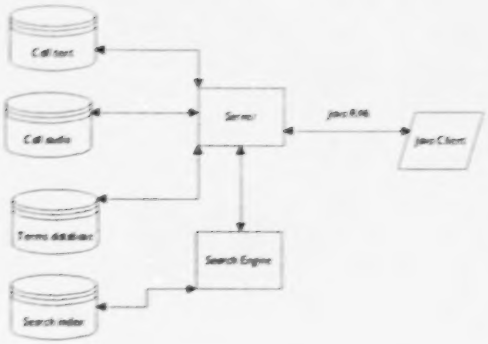

Figure 1- The Java client-server search system.

Figure 2 shows the Java client in action. After the user types in a query in the upper left entry field, the client sends the query to the server. The server returns a list of related terms from the Context thesaurus index, and a list of call titles from the document search index. Clicking on a particular call brings up a list of the terms in that call in the lower right list box.

Figure 2 – The Java search system, showing the query, the related terms, the call list and the terms in that call. The names and a few of the terms have been modified to preserve confidentiality.

7.1 Displaying the Call Documents

Once a user selects a specific call to investigate, it is important to present the call in the most meaningful way possible. Remember that most of the words in the call are inaccurate. However, we are fairly confident that the multi-word terms that Textract recognizes are likely to be correct. Thus, we constructed a system in which the server looked up the most salient terms in the transcript and marked them as word-objects in the data stream sent to the client. The client could then display these to best benefit the user.

We selected the terms in three stages until a large enough number of terms were found. This depended on the length and conversational density of the call transcription.

1. Select all multiword terms having an IQ >50.
2. If there are less than 10, select all single and multiword terms with an IQ >50.
3. If there are still fewer than 10, reduce the IQ limit to 30 and select all terms above that level.
4. Then, convert the word stream into individual token objects, and look for case insensitive matches within one character (to allow for plurals) of each successive word in the term.
5. Combine each located multiword term into a single token and mark it as salient.
6. Send the entire object stream to the client for display.

After some experimentation, we arrived at a client display in which the font of the terms not recognized as salient by Textract are shown as small as possible, so that they cannot easily be read, but only indicate the sequential position of words in the call.

We display the words that were found to be salient in a larger font, with a contrasting highlight color as shown in Figure 3. This client display was written in Java using the Java Swing JTextPane component, along with the DefaultStyledDocument, Highlighter and Highlight-Painter objects. We have described how to program these somewhat obscure components in another article [Cooper, 1999].

Figure 3- The document display and playback system

The display shown in Figure 3 illustrates a new way of looking at documents. Rather than just presenting a list of the salient terms in the document, this display shows the logical progress of the phone call and the points at which the important terms are discussed. These highlighted terms represent a kind of note taking on the actual phone call, and in a display represent a new kind of document genre, in which only the major terms are displayed to the users, but in which the viewer can deduce a general time line and grasp the progress of the call.

8. CALL PLAYBACK

Each of these calls was provided to us as digital files in PCM audio format. We converted these to the Sun au format using the copyaudio utility available from McGill University [McGill,

BEST COPY AVAILABLE

1999]. It was then possible to play back these files from the Java client server system. The Java client could request the call audio file from the server and receive it for playback on demand.

Since the JTextPane component can respond to clicks anywhere on its surface, and since we can calculate which word is clicked on, it is possible to construct a playback system in which the position of the word in the text stream can be recognized. Since word timing information was provided to us by the original speech engine output, we constructed each salient word object in the data stream to include this timing information. Then, the click on any word can be converted to its time offset in the call audio file. The time of a selected term is shown in the text field at the bottom of the window in Figure 3.

In order to play back the audio data, we used the Java Media Framework (JMF) to provide playback from a given time offset. The Java client JMF playback control requests the audio file from the web server and plays it when it arrives. While this is not "streaming audio," the time to download an entire audio file is short enough that the pause before the audio playback begins (even from the middle of a call) is quite acceptable when attached to a relatively high-speed network. Audio playback from a dialup connection was not successful.

The playback data window display is shown in Figure 4. It also contains the same JTextPane as well as the JMF audio player component.

8.1 Using JavaServer Pages for a Light Weight Display

While the Java client we described above is extremely powerful and flexible, it did not meet everyone's needs when restrictions of various browser types and Java RMI issues through firewalls were considered. Accordingly, we developed a second, lighter-weight client interface controlled by modified server code configured as a Java Bean that operated in conjunction with a JavaServer page [Pekowsky, 2000].

The JSP page makes Java calls to the Bean classes to obtain the context thesaurus and call information and fills two list boxes as shown in Figure 5.

Figure 4 – The JMF Playback Component combined with the JTextPane document display window.

It was also necessary to write a lightweight playback window that could work in this environment. Using DHTML and style sheets for highlighting we were able to devise a playback client which played the audio files using a Java 1.1 audio player. Highlighting of spans in DHTML was accomplished by statements like

```
<span class="yellow">
<a href="javascript:void 0" onClick=
"playSound(230.32)"> stock </a>
</span>
```

Clicking on a highlighted area starts the Java audio player at that point in the sound file. The playback client is shown in Figure 6.

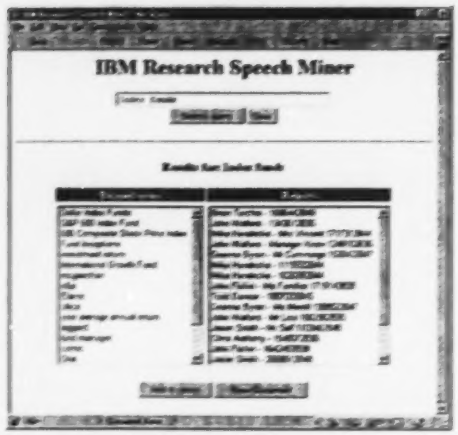

Figure 5 – A Lightweight client JavaServer page. Again, some names and terms have been slightly modified.

Figure 6 - A Lightweight playback client using DHTML. Some terms were modified as before.

9. TALKS AND REPORTS

Once we developed this process for indexing and a method of displaying the results that we found adequate, we applied the procedure to two new domains of discourse - consultant "notes from the field" and meetings. We accumulated data from two sources: dictation into portable recorders, and talks at a digitally recorded conference on knowledge management technology . Our two objectives were to find out if this system could be used to provide a way for e-Business consultants to make reports on customer engagements, and to find out if meetings could be summarized automatically.

9.1 Consultant Reports

We obtained several pocket digital recorders, (Olympus DS-150) and provided them to a series of IBM marketing and consulting people to record their reports in an experimental fashion. The idea behind this experiment was that consultants feel that their primary responsibility is to interact with customers and do not feel they are rewarded for writing reports on these interactions. Thus, it was

BEST COPY AVAILABLE

felt, important knowledge that might be of benefit to consultants in related engagements was lost.

The proposed scenario was that consultants would record these reports by speaking into the portable voice recorder after leaving the customer, and upload them to a web site we provided where we could produce a searchable database of these reports for other consultants.

Our technical findings were encouraging. The Olympus recorders are shipped with IBM ViaVoice and software for training the recognition system on your digitally recorded voice. Consultants, even with foreign accents, were readily recognized once they read a 100-sentence training script into the recorder. We generated an model for each consultant and processed their data. Following processing we created web pages much like those shown in Figure 6, and created a searchable index of the reports.

A segment of a typical report is shown below. (We note that "jowl one" actually refers to JavaOne.)

Activity report for July of the technology is marketing. One jowl one to tell that from June third to June ninth was made successful brief highlights an estimated 27 thousand people saw IBM sessions and DOS and jowl one. Of folks download trebled during that period over 7 hundred expanded IBM hospitality went. Over 5 z z taken from winnable PC initiative. 20 + articles

However, it developed that while this solution was technically quite feasible, the social aspects of this system were not at all what the sponsoring managers hoped. Even though both managers and consultants were quite enthusiastic about this system in theory, we received only a handful of reports over several months of the trial. We concluded that even though the consultants would like to have been able to browse a database of such reports, they had no impetus to help create this database, for the same reason they did not feel they had time to create written reports: it simply was not an important part of their job assignment.

So, even though we were easily able to create such a system, we found that it was quite difficult to interest the relevant practitioners in using it. There is a very important message here: technical systems will only be adopted by users if they are perceived not to impede their work flow, and if they are in some way rewarded for using them. In this case the rewards were too nebulous to justify the consultants' participation.

9.2 Conference Analysis

In our most recent experiment, we undertook the analysis of an internal marketing conference of knowledge management products. The conference was recorded on digital videotape. Each participant was given a lavaliere microphone to wear when they were the primary speaker.

We segmented the audio portion of he video tapes into individual presentations, converted the data into wave file format and used ViaVoice with a speaker-independent acoustic model to produce a transcript. For most speakers, the results were quite helpful. Figure 7 shows the results of one such presentation.

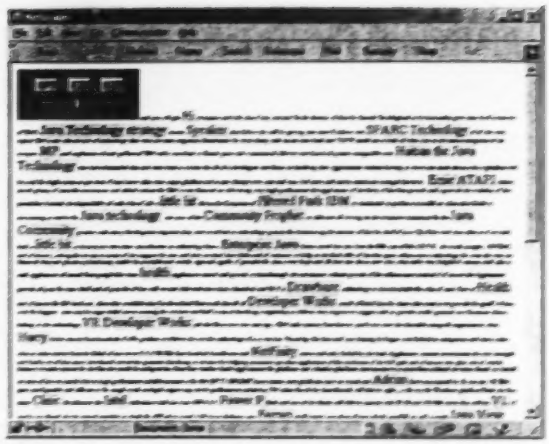

Figure 7 – A speech-recognized document from an internal knowledge management conference.

10. MEASURING ACCURACY USING SALIENT TERMS

If you look at the absolute word recognition accuracy of these transcriptions, you would find it to be as low as 20-30% in some cases, which at first seems a depressingly low number. This low accuracy is a result of the informal speech, incomplete sentences, regional accents and careless use of grammar and even pronunciation that frequently occur in telephone calls.

We selected the particularly problematic document described above for further analysis, and carefully transcribed the actual conversation, omitting stalling sounds such as "um" and "uh," and corrected for or omitted any overlapping dialog. We then ran this document through Textract as part of the collection of the remaining 522 documents, so that the same canonical forms could be developed in term recognition for this document.

After loading the results of Textract into our DB2 database, we compared the terms found in the recognized and manually transcribed documents. The results are shown in Table 2.

Table 2 –Recognition of Multiwords in a manually transcribed and automatically recognized document.

	manual	Automatic	%
Multiwords found in document	22	12	55
Multiwords less spurious finds	19	12	63
Multiwords where singles found	19	13.5	71

Of the several hundred terms found in both documents, we tabulated the multiword terms found in both. For the most part, these multiwords represent the high IQ salient concepts that are the highlighted terms in our display, and serve as a thumbnail outline of the conversation. There were 22 such terms in the manually transcribed document, where the opportunities for forming words correctly without intervening noise were greatest. Of those, 12 were found in the speech-recognized document.

However, once we eliminate multiwords caused by Textract "misfires" and the customer's name, which the recognition engine could not be expected to get, the correct rate rises to 12 out of 19.

9 BEST COPY AVAILABLE

Finally, if you give half-scores if all words of the multiword are found individually, we get 13.5 out of 19 or 71% accuracy in finding the salient concepts in the document.

Thus, we find, that even when very difficult conversations are subjected to speech recognition, the speech engine is quite capable of finding the preponderance of the salient terms in a document, and text mining systems like Textract are very capable of extracting these concepts and using them to provide note-like summaries of the major concepts in those conversations. In fact, even with extremely problematic recognition, caused by careless speakers and difficult regional accents, the ability of the speech engines to provide valuable document summaries remains very strong.

10.1 How Relevant Are the Terms?

Our final experiment deals with the question of the quality of the terms we are able to mine from speech transcripts. As outlined above, we start with raw speech data, process the output to detect sentence boundaries, and run the Textract text mining engine to find the main multi-word terms in the collection, and in each document.

The question we still needed to answer was whether the digital "notes" these documents represented were anything like the terms human subjects would find if asked to take notes on the same speech. If there is substantial overlap between the machine recognized multi-word terms and those found by the "note-takers," we can consider that the system is useful.

To test this hypothesis, we asked 10 subjects to watch an 18-minute segment of video from the meeting and take notes of the important concepts. We specifically asked for a note-taking style of lists of terms, to make sure that they used an approach that was similar to our automated system and similar to each other. In this experiment, we used the commercial version of IBM ViaVoice Millennium Edition, and an readily available TCL/TK toolkit for extracting the timing information from the ViaVoice data.

The results were quite encouraging. When we compared their results with the multi-words found by Textract on the voice-recognized transcript of the same 18-minute call, 17 keywords were found by all 10 human note-takers. These results are summarized in Table 3.

Table 3 – Terms found automatically and by 5 or more human note-takers in an 18-minute segment of video.

Phrase recognized by	Number of multi-words
Textract + 10 subjects	17
Textract + 9 subjects	25
Textract + 8 subjects	86
Textract + 7 subjects	91
Textract + 6 subjects	97
Textract + 5 subjects	102

We note that traditional measures of precision and recall are difficult to correlate with these data, because there is no agreed-upon way of determining the complete correct list of terms: it is subjective both from the human subject and the computational point of view.

From these preliminary experiments, we conclude that these keyword-highlighted summary representations of speech-recognition output created by our system are sufficiently accurate to represent a useful set of meeting notes that can be searched and displayed for use in playback of such presentations.

11. RESULTS AND DISCUSSION

In these experiments, we discovered that while it is not yet possible to produce word-accurate transcripts of informal conversations, reports and meetings, it is still possible to provide extremely useful information regarding the content of this voice data. Rather than regarding the recognized text as a transcript, it is more useful to consider it a beginning of a set of "notes" on the event such as a party might take down to remind themselves later of what was said.

We found that by recognizing the salient terms in the conversation and providing a search system for searching the call or report archive, we can provide supervisors and other consultants with a way of looking into what information was discussed, and a way of playing back the interesting portions of conversations returned from such a search.

Finally, we found that while the word recognition accuracy of these transcripts was in many cases fairly low, the salient term accuracy was quite high and made these searchable summaries extremely useful.

We note here that while the Textract tool we used here is an internal research prototype, there is a product version available and that commercial products from other vendors may also be used. In general, we have found that the Textract tools is more efficient in aggregating variant forms of terms than most of the current commercial systems, but for this particular application, they might well be considered roughly equivalent.

12. ACKNOWLEDGMENTS

We'd like to thank Zunaid Kazi for his work in developing the unnamed relations and context thesaurus code, Eric Brown for recording the conference on digital videotape, Branimir Boguraev for helpful suggestions which led to the development of the client display, Yael Ravin for modifications to Textract to make name detection more reliable in this environment, Paul Kaye for writing the media player interface, Bob Mack for helpful general suggestions, Salim Roukos for helpful conversations, Harry Kolar for thoughtful analysis of the results, Michael Hebenberger for his work in developing the project, and Roy Byrd and Alan Marwick for their technical support.

13. BIBLIOGRAPHY

[1] Brandow, Ron, Karl Mitze, and Lisa Rau, 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31, No. 5, 675-685.

[2] Byrd, R.J. and Ravin, Y. Identifying and Extracting Relations in Text. *Proceeedings of NLDB 99*, Klagenfurt, Austria.

[3] Buckley, C., Singhal, A., Mira, M & Salton, G. (1996) "New Retrieval Approaches Using SMART:TREC4. In Harman, D, editor, Proceedings of the TREC 4 Conference, National Institute of Standards and Technology Special Publication.

[4] Chen, Scott Shaobing, M.J.F. Gales, P.S. Gopalakrishnan,

R.A. Gopinath, H. Printz, D. Kanevsky, P. Olsen, L. Polymenakos, IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub-4 English Evaluation in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998

[5] Cooper, J.W. and Prager, John M, *Anti-Serendipity: Finding Useless Documents and Similar Documents*, Proceedings of HICSS-33, Maui, HI, January, 2000.

[6] Cooper, J.W., "Colorful Language," JavaPro 4(1), 44, 2000.

[7] Cooper, James W. and Byrd, Roy J. "Lexical Navigation: Visually Prompted Query Expansion and Refinement." Proceedings of DIGLIB97, Philadelphia, PA, July, 1997.

[8] Cooper, James W. and Byrd, Roy J., OBIWAN – "A Visual Interface for Prompted Query Refinement," Proceedings of HICSS-31, Kona, Hawaii, 1998.

[9] Dharanipragada, S. and S. Roukos, Experimental Results in Audio Indexing in Proceedings of the DARPA Speech Recognition Workshop, 1997

[10] Fowler, Richard H., Wilson, Bradley A., and Fowler, Wendy A.L. "Information Navigator: An information system using networks for display and retrieval." Report NAG9-551, No.92-1. Department of Computer Science, University of Texas, Pan American, Edinburg, TX.

[11] Johnson, S.E., Jourlin, P., Spark-Jones, K. and Woodland, P.C. *Spoken Document Retrieval for TREC-8 at Cambridge University*, Proceedings of TREC-8, 1999.

[12] McGill University FTP Archive. See ftp.TSP.ECE.McGill.CA/pub/Afsp

[13] Neff, Mary S. and Cooper, James W. 1999a. Document Summarization for Active Markup, *in Proceedings of the 32nd Hawaii International Conference on System Sciences*, Wailea, HI, January, 1999.

[14] Pekowsky, Lame. *JavaServer Pages*, Addison-Wesley, Boston, MA, 2000.

[15] Prager, John M., Linguini: Recognition of Language in Digital Documents, *in Proceedings of the 32nd Hawaii International Conference on System Sciences*, Wailea, HI, January, 1999.

[16] Schatz, Bruce R, Johnson, Eric H, Cochrane, Pauline A and Chen, Hsinchun, "Interactive Term Suggestion for Users of Digital Libraries." *ACM Digital Library Conference, 1996.*

[17] Viswanathan, M, H.S.M. Beigi, S. Dharanipragada, and A. Tritschler, "Retrieval from Spoken Documents Using Content And Speaker Information," Proceedings, International Conference on Document Analysis and Retrieval (ICDAR99), Bangalore, India, 1999, pp. 567--572.

[18] Woodland, P.C., Johnson, S.E., Jourlin, P. and Sparck-Jones, K. *Effects of Out of Vocabulary Words in Spoken Document Retrieval*. Proceedings of SIGIR 2000, Athens, Greece, July, 2000.

[19] Xu, Jinxi and Croft, W. Bruce. "Query Expansion Using Local and Global Document Analysis," *Proceedings of the 19th Annual ACM-SIGIR Conference*, 1996, pp. 4-11